# Lab 10

## B)

4. There are 3 classes (0,1,2) which are evenly distributed, and each contain 4 descriptive features.

## D)

1. The best performing classifier was k-nearest neighbors for both 80/20 and 70/30 splits and was able to reach an accuracy of ~93% ~95% respectively. This makes sense as the knn algorithm uses the actual data to build its predictions and is a good match for a simple dataset without inconsistencies, like the iris dataset. The decision tree performed about as well as the linear regression and support vector machine models, which is interesting, as typically SVM will outperform both traditional linear regression and decision tree algorithms.

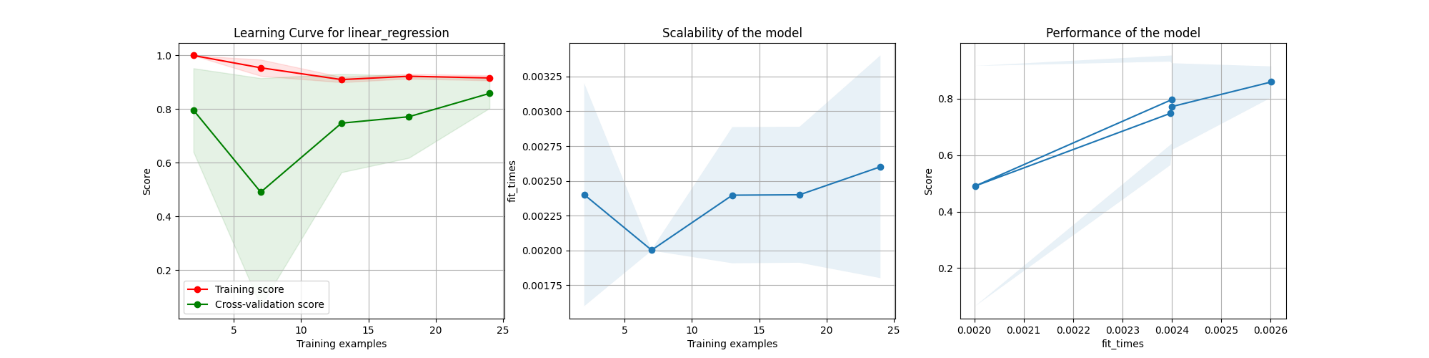
2. In adjusting the testing split from the initial 80/20 to 70/30, we actually see increased performance across each of the estimator classes. This means we were most likely overfitting the base dataset. We can tell this because by examining less training examples, we were performing better on unseen data. This meant that we were picking up on patterns present in our training set that would not scale to the rest of the data, resulting in bias.

Below are the learning curves for each of the estimators for the 80/20 and 70/30 splits, as well as their calculated score.

### LinReg

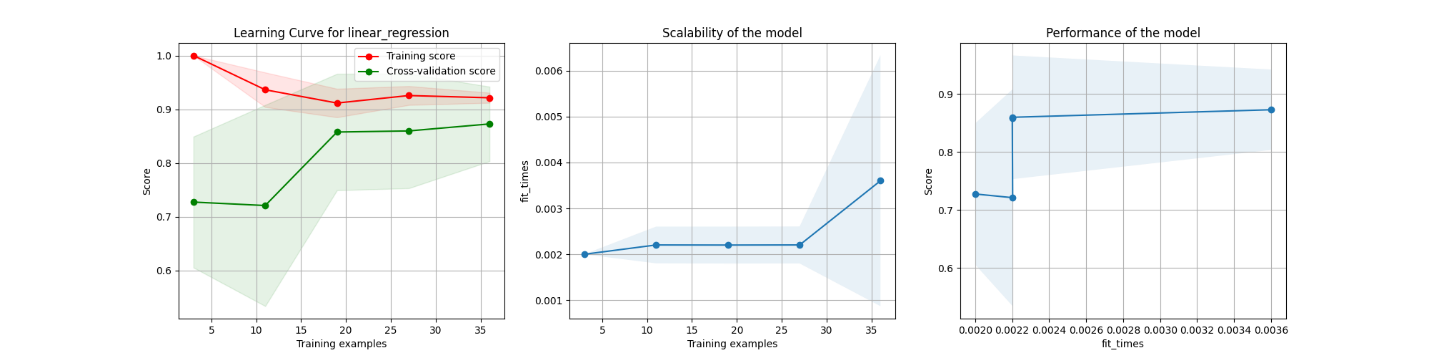
#### 80/20:

Score: 89.8%



#### 70/30:

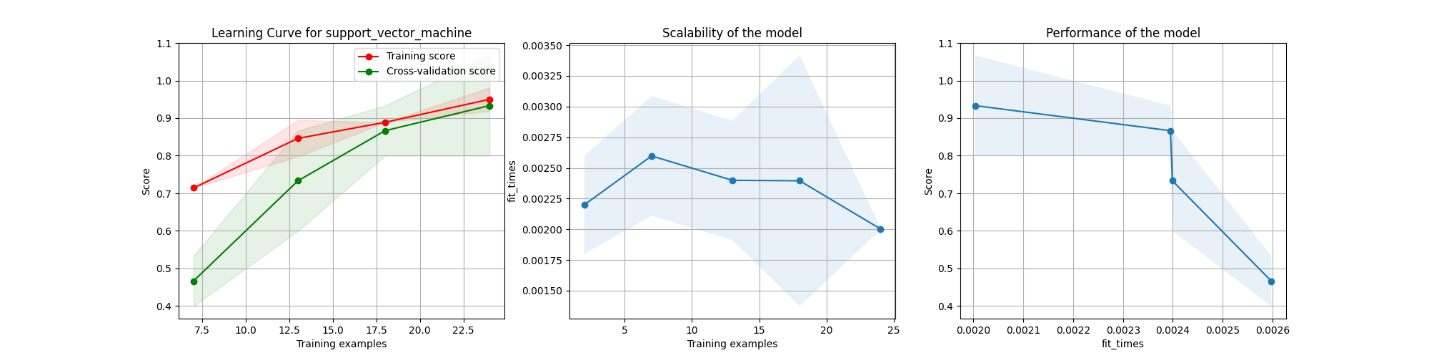
Score: 91.3%



### SVM

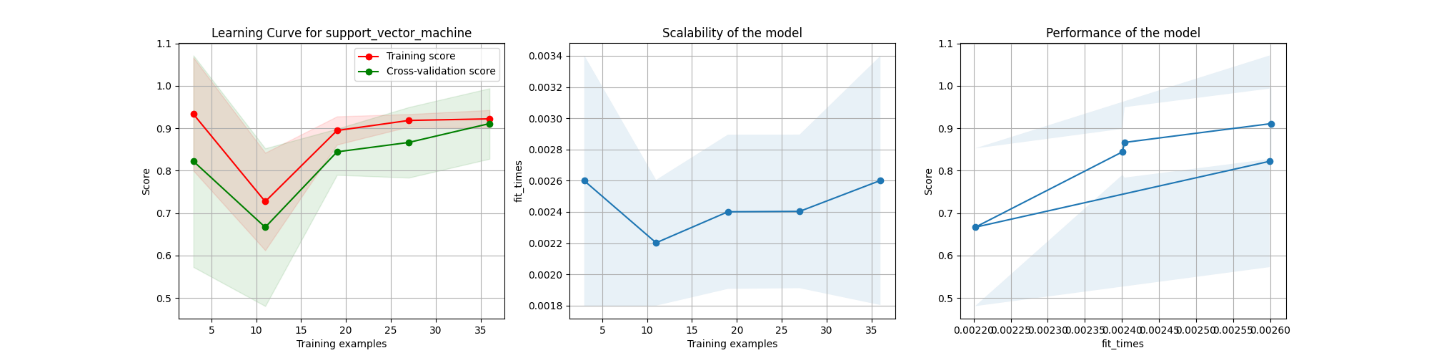
#### 80/20:

Score: 90%



#### 70/30:

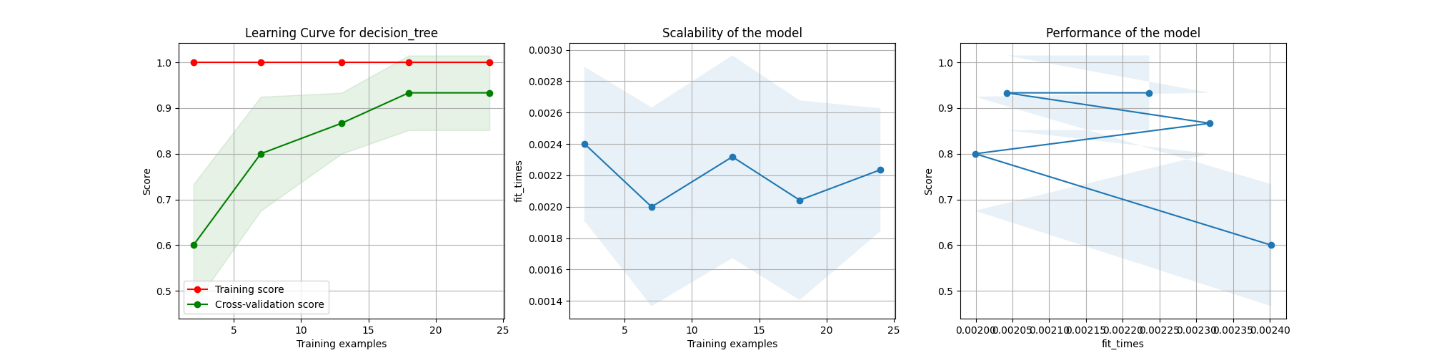
Score: 93.3%



### DT

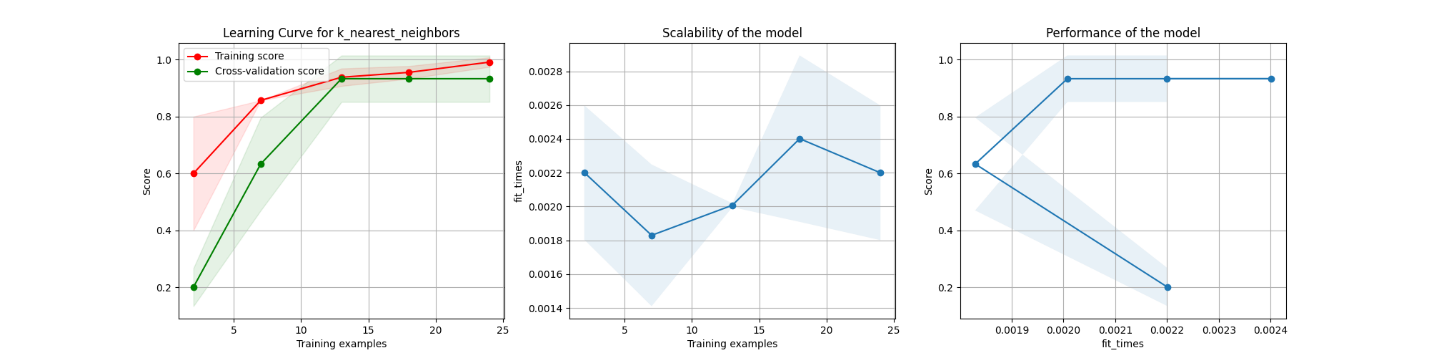
#### 80/20:

Score: 90%



#### 70/30:

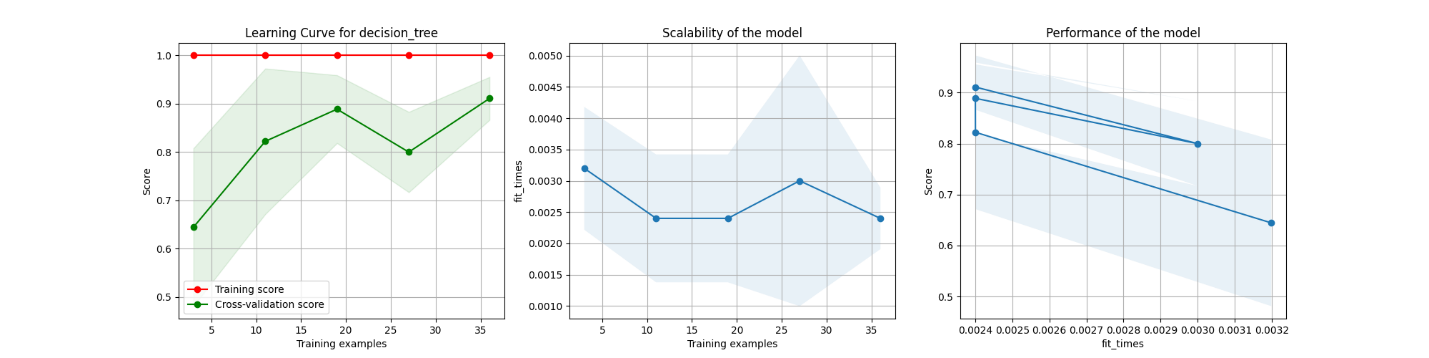
Score: 91.1%



### KNN

#### 80/20:

Score: 93.3%



#### 70/30:

Score: 95.6%

